(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平8－328760

(43)公開日　平成８年(1996)12月13日

(51)Int.Cl.⁶	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　3/06	５４０		Ｇ０６Ｆ　3/06	５４０
13/10	３４０	7368－5Ｅ	13/10	３４０Ａ

審査請求　未請求　請求項の数14　ＦＤ　（全 21 頁）

(21)出願番号　特願平7－158370

(22)出願日　平成７年(1995)６月１日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　角田　仁
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内

(72)発明者　大山　光男
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内

(72)発明者　高本　良史
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内

(74)代理人　弁理士　矢島　保夫

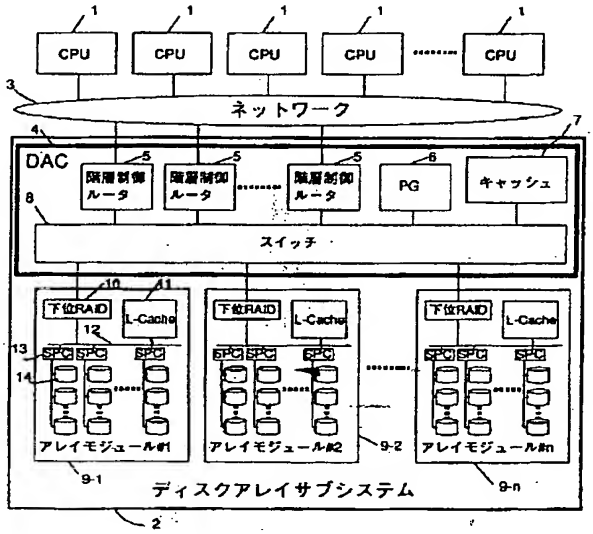
最終頁に続く

(54)【発明の名称】　ディスクアレイ装置

(57)【要約】

【目的】複数のディスクアレイモジュール間でディスクアレイ制御を実現する場合において、ディスクアレイ制御装置の基板上の実装の制約を排除し、制御を簡単にして転送速度を向上させることのできるディスクアレイ装置を提供することを目的とする。

【構成】複数のディスクアレイモジュール間でディスクアレイ制御を実現する場合、従来のように複数のディスクアレイ制御装置をマザーボードに接続し、このマザーボード内の配線によるバスでこれらの複数のディスクアレイ制御装置間を制御するのではなく、並列動作が可能なクロスバ方式などのスイッチ手段により複数のディスクアレイモジュールを接続し、このスイッチのルーティング制御により複数のディスクアレイモジュール間によるディスクアレイ制御を実現する。



1

【特許請求の範囲】

【請求項１】上位装置に接続され、複数台のディスクアレイモジュール間でディスクアレイ制御を行うディスクアレイ装置であって、

上位装置から発行された読み出しまたは書き込み要求を受け付けるルータと、

各々が独立したディスクアレイ装置として内部でディスクアレイ制御を行っている複数台のディスクアレイモジュールと、

上記ルータ、および上記複数台のディスクアレイモジュールを各ポートに接続するとともに、それら各ポート間の接続を行うスイッチ手段とを備え、

上記ルータにより上記スイッチ手段の各ポート間の接続を制御することにより、上記複数台のディスクアレイモジュール間でディスクアレイ制御を行うことを特徴とするディスクアレイ装置。

【請求項２】上位装置に接続され、複数台のディスクアレイモジュール間でディスクアレイ制御を行うディスクアレイ装置であって、

上位装置から発行された読み出しまたは書き込み要求を受け付けるルータと、

パリティを生成するためのパリティ生成手段と、

各々が独立したディスクアレイ装置として内部でディスクアレイ制御を行っている複数台のディスクアレイモジュールと、

上記ルータ、上記パリティ生成手段、および上記複数台のディスクアレイモジュールを各ポートに接続するとともに、それら各ポート間の接続を行うスイッチ手段とを備え、

上記ルータにより上記スイッチ手段の各ポート間の接続を制御することにより、上記複数台のディスクアレイモジュール間でディスクアレイ制御を行うことを特徴とするディスクアレイ装置。

【請求項３】さらに、前記スイッチ手段のポートにキャッシュメモリを接続し、前記上位装置から発行された読み出しまたは書き込み要求に対して読み出しまたは書き込みを行うべきディスクアレイモジュールが接続されているポートを認識するためのルーティングテーブルを前記キャッシュメモリに記憶しておき、

前記ルータは、前記ルーティングテーブルを用いて、前記上位装置から発行された読み出しまたは書き込み要求に対して読み出しまたは書き込みを行うべきディスクアレイモジュールが接続されているポートを認識し、該ポートに接続されているディスクアレイモジュール間でディスクアレイ制御を行う請求項１または２に記載のディスクアレイ装置。

【請求項４】前記ルーティングテーブルを、前記キャッシュメモリ内に設ける代わりに、前記ルータから前記スイッチ手段を介さずに直接アクセスできるメモリ内に設けた請求項３に記載のディスクアレイ装置。

2

【請求項５】前記ルータは、前記上位装置からの書き込み要求を受け付けたとき、該書き込みデータを前記スイッチ手段を介して前記パリティ生成手段に転送し、前記パリティ生成手段において書き込みデータの分割およびパリティの生成を行い、該分割したデータおよび生成したパリティを、前記ルーティングテーブルを用いて認識された複数のディスクアレイモジュールに対してそれぞれ転送して書き込む請求項３または４に記載のディスクアレイ装置。

【請求項６】前記ルータは、前記上位装置からの読み出し要求を受け付けたとき、前記ルーティングテーブルを用いて、読み出すべき分割されたデータが格納されている複数のディスクアレイモジュールを認識し、前記スイッチ手段を介して該複数のディスクアレイモジュールと前記パリティ生成手段とを接続し、該複数のディスクアレイモジュールから前記パリティ生成手段へと分割されたデータを読み出し、前記パリティ生成手段は該読み出された分割されたデータを結合し、該結合したデータを前記パリティ生成手段から前記スイッチ手段および前記ルータを介して前記上位装置に転送する請求項３または４に記載のディスクアレイ装置。

【請求項７】前記ルータは、前記上位装置からの読み出し要求を受け付けた場合、読み出すべき分割されたデータが格納されている複数のディスクアレイモジュールのうちの何れかに障害が発生していたときは、前記ルーティングテーブルを用いて、読み出すべき分割されたデータ（障害が発生しているディスクアレイモジュールに格納されているデータを除く）が格納されている複数のディスクアレイモジュールおよびパリティが格納されているディスクアレイモジュールを認識し、前記スイッチ手段を介してそれらのディスクアレイモジュールと前記パリティ生成手段とを接続し、それらのディスクアレイモジュールから前記パリティ生成手段へと分割されたデータおよびパリティを読み出し、前記パリティ生成手段は該読み出された分割されたデータおよびパリティから障害が発生したディスクアレイモジュール内のデータを回復し、該回復したデータと読み出された分割されたデータを結合し、該結合したデータを前記パリティ生成手段から前記スイッチ手段および前記ルータを介して前記上位装置に転送する請求項６に記載のディスクアレイ装置。

【請求項８】前記ルータは、前記上位装置からの書き込み要求を受け付けたとき、前記ルーティングテーブルを用いて、該書き込みデータに対応する旧データおよび旧パリティが格納されているディスクアレイモジュールを認識し、該旧データおよび旧パリティを前記スイッチ手段を介して前記パリティ生成手段へと読み出し、前記パリティ生成手段は、読み出された旧データおよび旧パリティと書き込みデータとを用いて新パリティを生成し、該書き込みデータおよび生成した新パリティを、旧デー

タおよび旧パリティを読み出したディスクアレイモジュールに転送して書き込む請求項３または４に記載のディスクアレイ装置。

【請求項９】前記ルータは、前記上位装置からの読み出し要求を受け付けたとき、前記ルーティングテーブルを用いて、読み出すべきデータが格納されているディスクアレイモジュールを認識し、該ディスクアレイモジュールから読み出したデータを、前記スイッチ手段および前記ルータを介して前記上位装置に転送する請求項３または４に記載のディスクアレイ装置。

【請求項１０】前記ルータは、前記上位装置からの読み出し要求を受け付けた場合、読み出したいデータが格納されているディスクアレイモジュールに障害が発生していたときは、前記ルーティングテーブルを用いて、当該読み出したいデータが作成に関与したパリティが格納されているディスクアレイモジュールおよび該パリティの作成に関与した当該読み出したいデータ以外のデータが格納されているディスクアレイモジュールを認識し、前記スイッチ手段を介してそれらのディスクアレイモジュールと前記パリティ生成手段とを接続し、それらのディスクアレイモジュールから前記パリティ生成手段へとパリティおよび該パリティの作成に関与したデータを読み出し、前記パリティ生成手段は該読み出されたパリティおよび該パリティの作成に関与したデータから障害が発生したディスクアレイモジュール内のデータを回復し、該回復したデータを前記パリティ生成手段から前記スイッチ手段および前記ルータを介して前記上位装置に転送する請求項９に記載のディスクアレイ装置。

【請求項１１】前記上位装置が前記ルータに直接接続されており、前記ルータは、前記上位装置からの読み出しまたは書き込み要求を直接受け付ける請求項１または２に記載のディスクアレイ装置。

【請求項１２】前記上位装置が前記スイッチ手段のポートに直接接続されており、前記ルータは、前記スイッチ手段を介して前記上位装置からの読み出しまたは書き込み要求を受け付ける請求項１または２に記載のディスクアレイ装置。

【請求項１３】前記複数台のディスクアレイモジュールのうち任意の数のディスクアレイモジュールを、ディスクアレイ装置の本体とは別の筐体とした請求項１または２に記載のディスクアレイ装置。

【請求項１４】請求項１または２のディスクアレイ装置のうち、複数台のディスクアレイモジュール以外の部分をディスクアレイ制御装置と呼ぶとき、複数のディスクアレイ制御装置で共通の複数台のディスクアレイモジュールを共有する請求項１または２に記載のディスクアレイ装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、コンピュータシステムにおける２次記憶装置に関し、特に高性能な入出力動作を可能とするディスクアレイ装置に関する。

【０００２】

【従来の技術】現在のコンピュータシステムにおいては、ＣＰＵ（中央処理装置）などの上位側が必要とするデータは２次記憶装置に格納され、ＣＰＵなどが必要とするときに応じて２次記憶装置に対してデータの書き込みおよび読み出しを行っている。この２次記憶装置としては、一般に不揮発な記憶媒体が使用され、代表的なものとして磁気ディスク装置（以下、ドライブとする）や、光ディスクなどがあげられる。

【０００３】近年高度情報化に伴い、コンピュータシステムにおいて、この種の２次記憶装置の高性能化が要求されてきた。その一つの解として、多数の比較的容量の小さなドライブにより構成されるディスクアレイが考えられている。

【０００４】公知の文献として、「D.Patterson、G.Gibson、and R.H.Kartz;A Case for Redundant Arrays of Inexpensive Disks(RAID)、in ACM SIGMOD Conference、Chicago、IL、(June1988)」がある。この文献においては、全く同じデータを別のドライブに二重化して格納するミラーリングを行うディスクアレイ（レベル１）と、データを分割して並列に処理を行うディスクアレイ（レベル３）と、データを分散して独立に扱うディスクアレイ（レベル４、５）について、その性能および信頼性の検討結果が報告されている。レベル４は、レベル５において論理グループを構成するドライブに分散しているパリティを、１台のパリティのみを格納するドライブにまとめたものである。現在この論文に書かれている方式が、最も一般的なディスクアレイと考えられている。

【０００５】ここで、レベル３、レベル４、およびレベル５のそれぞれについて、簡単に説明しておく。

【０００６】まず、レベル３のディスクアレイについて簡単に説明する。ディスクアレイに格納するデータ＃１として、例えば「００１０１０１０１１・・・・」を想定し、このデータ＃１とパリティを格納するためのディスクとしてディスク＃１～＃５が設けられているとする。レベル３では、ディスク＃１にデータ＃１の第１ビット‘０’を、ディスク＃２にその次の第２ビット‘０’を、ディスク＃３に第３ビット‘１’を、ディスク＃４に第４ビット‘０’を、順次格納し、格納された‘００１０’に対するパリティをディスク＃５に格納する。そして、次に同様にして、引き続くビット‘１’，‘０’，‘１’，‘０’を順次ディスク＃１～＃４に格納し、そのパリティを＃５に格納してゆく。

【０００７】レベル４では、データとパリティを格納するためのディスクとしてディスク＃１～＃５が設けられた場合、データ＃１，＃５，・・・がディスク＃１に、データ＃２，＃６，・・・がディスク＃２に、データ＃３，＃７，・・・がディスク＃３に、データ＃４，＃

８，・・・がディスク＃４に、それぞれ格納される。そして、例えば、データ＃１が「０１・・・」、データ＃２が「００・・・」、データ＃３が「１１・・・」、データ＃４が「００・・・」であるとすると、各データの先頭ビットを並べた‘００１０’に対するパリティをパリティ専用として指定されたディスク＃５の先頭ビットととして格納し、次に各データの２番目のビット‘１０１０’に対するパリティをディスク＃５の２番目のビットととして格納し、以下同様にしていく。そして、データ＃５～＃８のデータ組に対するパリティデータを、ディスク＃５に２番目のパリティデータとして、格納するようにしてゆく。

【０００８】レベル５は、レベル４のようにパリティ専用のディスクを決めず、データ＃１をディスク＃１に、データ＃２をディスク＃２に、データ＃３をディスク＃３に、データ４をディスク＃４にそれぞれ格納し、データ＃１～＃４のデータ組に対するパリティデータをディスク＃５に格納し、次いで、データ＃５をディスク＃２に、データ＃６をディスク＃３に、データ＃７をディスク＃４に、データ８をディスク＃５にそれぞれ格納し、データ＃５～＃８のデータ組に対するパリティデータをディスク＃１に格納し、次いで、データ＃９をディスク＃１に、データ＃１０をディスク＃３に、データ＃１１をディスク＃４に、データ１２をディスク＃５にそれぞれ格納し、データ＃９～＃１２のデータ組に対するパリティデータをディスク＃２に格納する、とようにしてゆく。要するに、パリティデータを分散して格納するものである。

【０００９】上記文献に記載されたタイプのディスクアレイでは、大型大容量のドライブを、多数の比較的容量の小さなドライブで構成し、データを分散して格納するため、読み出し／書き込み要求が増加してもディスクアレイの複数のドライブで分散して処理することが可能となり、読み出し／書き込み要求が待たされることが減少する。

【００１０】次に、ディスクアレイにおけるパリティについて説明する。ディスクアレイは従来の大容量のドライブを、比較的容量の小さな多数のドライブで構成するため、部品点数が増加し障害が発生する確率が高くなる。このため、ディスクアレイでは、上述したようにパリティを用意している。パリティを用意することにより、データを格納したドライブに障害が発生した場合でもその障害ドライブ内のデータを復元することが可能となる。ディスクアレイでは、データからパリティを作成しデータと同様にドライブに格納しておく。このとき、パリティは、パリティの作成に関与したデータとは別のドライブに格納される。

【００１１】特開平６－１１９１２０号公報は、ディスクアレイにおけるデータの更新時に、キャッシュメモリ内に当該データが存在する場合は、キャッシュメモリ内の当該データを更新し、所定回数の更新が行われた場合、ディスクアレイ内の冗長データ（パリティ）を更新する方法について開示している。この特開平６－１１９１２０号公報では、ディスクアレイ制御装置の構成例が図２および図３に示されている。ここで示されている構成例は、現在最も一般的な構成方法と考えられる。具体的には、ディスクアレイ制御装置には制御手段として動作するＭＰＵ（マイクロプロセッサ）が設けられ、ＭＰＵからの内部バスに対し、処理プログラムを格納したＲＯＭ（リード・オンリ・メモリ）、制御記憶等として用いられるＲＡＭ（ランダム・アクセス・メモリ）、キャッシュ制御部を介して接続したキャッシュメモリ、およびデータ転送バッファが設けられる。このように、ディスクアレイ制御装置内において、ＭＰＵ、ＲＯＭ、ＲＡＭ、およびキャッシュ制御部は、これらが共有で使用する内部バスにより相互に接続されている。

【００１２】また、別の公知例としては特開平６－２４２８８７号公報がある。この公知例では、ＣＰＵとディスクアレイ（ディスクアレイコントローラ＋ドライブ）とをネットワークスイッチにより接続することで、ディスクアレイのＩ／Ｏ帯域幅とＣＰＵのＩ／Ｏ性能を最適に整合させる方法について開示している。つまり、高性能かつ自由度の高いネットワークスイッチを介してＣＰＵとディスクアレイとを接続することで、接続の自由度を向上させ、ネックを解消するようにしている。

【００１３】

【発明が解決しようとする課題】現在のディスクアレイ制御装置は、特開平６－１１９１２０号公報に示されているように、制御手段として動作するＭＰＵが設けられ、処理プログラムを格納したＲＯＭ、制御記憶等として用いられるＲＡＭ、キャッシュ制御部を介して接続したキャッシュメモリ、およびデータ転送バッファは、内部バスを介してＭＰＵに接続されている。このように、ディスクアレイ制御装置内では、ＭＰＵにおけるすべての制御が内部バスを介して行われる。この内部バスは、実質的にはＭＰＵ、ＲＯＭ、ＲＡＭ、およびキャッシュ等が実装されている基板上の配線である。この、基板上の配線による内部バスの性能は、内部バスの転送速度（ＭＢ／ｓ）と内部バスのバス幅（Ｂ）との積で決定される。しかし、基板が大きくなり配線長が長くなると、転送速度を向上させることが困難になる。特に、今後複数のディスクアレイ制御装置間でディスクアレイ制御を実現しようとした場合、複数のディスクアレイ制御装置が接続されるマザーボードはかなり大きくなるため、転送速度を向上させることはさらに困難になる。

【００１４】また、内部バスのバス幅を広げると基板内での実装やコネクタが大きくなるため、実装上の問題が生じる。

【００１５】以上より、今後ディスクアレイ制御装置の性能を向上させるためには、このような基板内での配線

を使用する内部バスが性能のネックになるという問題を解決しなければならない。

【００１６】本発明は、ディスクアレイ制御装置の基板上の実装の制約を排除し、制御を簡単にして転送速度を向上させることのできるディスクアレイ装置を提供することを目的とする。

【００１７】

【課題を解決するための手段】上記目的を達成するため、請求項１に記載の発明は、上位装置に接続され、複数台のディスクアレイモジュール間でディスクアレイ制御を行うディスクアレイ装置であって、上位装置から発行された読み出しまたは書き込み要求を受け付けるルータと、各々が独立したディスクアレイ装置として内部でディスクアレイ制御を行っている複数台のディスクアレイモジュールと、上記ルータ、および上記複数台のディスクアレイモジュールを各ポートに接続するとともに、それら各ポート間の接続を行うスイッチ手段とを備え、上記ルータにより上記スイッチ手段の各ポート間の接続を制御することにより、上記複数台のディスクアレイモジュール間でディスクアレイ制御を行うことを特徴とする。

【００１８】また、請求項２に記載の発明は、請求項１に記載の発明の構成に加えてパリティ生成手段をもスイッチ手段に接続するようにして、ＲＡＩＤレベル４、５にも適用できる点を明らかにしたものである。

【００１９】スイッチ手段の各ポート間の接続を制御することにより複数台のディスクアレイモジュール間でディスクアレイ制御を行うため、上位装置から発行された読み出しまたは書き込み要求に対して読み出しまたは書き込みを行うべきディスクアレイモジュールが接続されているポートを認識するためのルーティングテーブルを用いるとよい。ルーティングテーブルは、キャッシュメモリ内に設けてもよいし、スイッチ手段を介さずに直接アクセスできるメモリ内に設けてもよい。

【００２０】請求項５に記載の発明は、請求項３または４に記載のディスクアレイ装置においてＲＡＩＤレベル３のディスクアレイ制御を行う際の書き込み要求に対する処理を明らかにしたものであり、前記上位装置からの書き込み要求を受け付けたとき、該書き込みデータを前記スイッチ手段を介して前記パリティ生成手段に転送し、前記パリティ生成手段において書き込みデータの分割およびパリティの生成を行い、該分割したデータおよび生成したパリティを、前記ルーティングテーブルを用いて認識された複数のディスクアレイモジュールに対してそれぞれ転送して書き込むことを特徴とする。

【００２１】請求項６に記載の発明は、請求項３または４に記載のディスクアレイ装置においてＲＡＩＤレベル３のディスクアレイ制御を行う際の読み出し要求に対する処理を明らかにしたものであり、前記上位装置からの読み出し要求を受け付けたとき、前記ルーティングテーブルを用いて、読み出すべき分割されたデータが格納されている複数のディスクアレイモジュールを認識し、前記スイッチ手段を介して該複数のディスクアレイモジュールと前記パリティ生成手段とを接続し、該複数のディスクアレイモジュールから前記パリティ生成手段へと分割されたデータを読み出し、前記パリティ生成手段は該読み出された分割されたデータを結合し、該結合したデータを前記パリティ生成手段から前記スイッチ手段および前記ルータを介して前記上位装置に転送することを特徴とする。

【００２２】請求項７に記載の発明は、さらにＲＡＩＤレベル３のディスクアレイ制御を行う際の読み出し要求で、読み出すべき分割されたデータが格納されている複数のディスクアレイモジュールのうちの何れかに障害が発生していたときの処理を明らかにしたものである。すなわち、前記ルータは、前記上位装置からの読み出し要求を受け付けた場合、読み出すべき分割されたデータが格納されている複数のディスクアレイモジュールのうちの何れかに障害が発生していたときは、前記ルーティングテーブルを用いて、読み出すべき分割されたデータ（障害が発生しているディスクアレイモジュールに格納されているデータを除く）が格納されている複数のディスクアレイモジュールおよびパリティが格納されているディスクアレイモジュールを認識し、前記スイッチ手段を介してそれらのディスクアレイモジュールと前記パリティ生成手段とを接続し、それらのディスクアレイモジュールから前記パリティ生成手段へと分割されたデータおよびパリティを読み出し、前記パリティ生成手段は該読み出された分割されたデータおよびパリティから障害が発生したディスクアレイモジュール内のデータを回復し、該回復したデータと読み出された分割されたデータを結合し、該結合したデータを前記パリティ生成手段から前記スイッチ手段および前記ルータを介して前記上位装置に転送することを特徴とする。

【００２３】請求項８に記載の発明は、請求項３または４に記載のディスクアレイ装置においてＲＡＩＤレベル４、５のディスクアレイ制御を行う際の書き込み要求に対する処理を明らかにしたものであり、前記ルータは、前記上位装置からの書き込み要求を受け付けたとき、前記ルーティングテーブルを用いて、該書き込みデータに対応する旧データおよび旧パリティが格納されているディスクアレイモジュールを認識し、該旧データおよび旧パリティを前記スイッチ手段を介して前記パリティ生成手段へと読み出し、前記パリティ生成手段は、読み出された旧データおよび旧パリティと書き込みデータとを用いて新パリティを生成し、該書き込みデータおよび生成した新パリティを、旧データおよび旧パリティを読み出したディスクアレイモジュールに転送して書き込むことを特徴とする。

【００２４】請求項９に記載の発明は、請求項３または

4に記載のディスクアレイ装置においてRAIDレベル4、5のディスクアレイ制御を行う際の読み出し要求に対する処理を明らかにしたものであり、前記ルータは、前記上位装置からの読み出し要求を受け付けたとき、前記ルーティングテーブルを用いて、読み出すべきデータが格納されているディスクアレイモジュールを認識し、該ディスクアレイモジュールから読み出したデータを、前記スイッチ手段および前記ルータを介して前記上位装置に転送することを特徴とする。

【0025】請求項10に記載の発明は、さらにRAIDレベル4、5のディスクアレイ制御を行う際の読み出し要求で、読み出すべきデータが格納されているディスクアレイモジュールに障害が発生していたときの処理を明らかにしたものである。すなわち、前記ルータは、前記上位装置からの読み出し要求を受け付けた場合、読み出したいデータが格納されているディスクアレイモジュールに障害が発生していたときは、前記ルーティングテーブルを用いて、当該読み出したいデータが作成に関与したパリティが格納されているディスクアレイモジュールおよび該パリティの作成に関与した当該読み出したいデータ以外のデータが格納されているディスクアレイモジュールを認識し、前記スイッチ手段を介してそれらのディスクアレイモジュールと前記パリティ生成手段とを接続し、それらのディスクアレイモジュールから前記パリティ生成手段へとパリティおよび該パリティの作成に関与したデータを読み出し、前記パリティ生成手段は該読み出されたパリティおよび該パリティの作成に関与したデータから障害が発生したディスクアレイモジュール内のデータを回復し、該回復したデータを前記パリティ生成手段から前記スイッチ手段および前記ルータを介して前記上位装置に転送することを特徴とする。

【0026】本発明では、上位装置からの読み出しまたは書き込み要求はルータにより受け付けられ、ルータがスイッチ手段を制御してRAID制御を行う。この場合、上位装置とルータとを直接接続してもよいし、スイッチ手段を介して接続してもよい。

【0027】複数台のディスクアレイモジュールのうち任意の数のディスクアレイモジュールを、ディスクアレイ装置の本体とは別の筐体としてもよい。また、複数台のディスクアレイモジュール以外の部分をディスクアレイ制御装置と呼ぶとき、複数のディスクアレイ制御装置で共通の複数台のディスクアレイモジュールを共有するようにしてもよい。

【0028】

【作用】本発明によれば、複数のディスクアレイ制御装置間でディスクアレイ制御を実現する場合、従来のように複数のディスクアレイ制御装置をマザーボードに接続し、このマザーボード内の配線によるバスでこれらの複数のディスクアレイ制御装置間を制御するのではなく、例えば、並列動作が可能なクロスバ方式等のスイッチなどのスイッチ手段により複数のディスクアレイモジュールを接続し、このスイッチ手段のルーティング制御により複数のディスクアレイモジュール間によるディスクアレイ制御を実現する。

【0029】複数のディスクアレイモジュール間によるディスクアレイ制御をクロスバ方式などのスイッチ手段のルーティング制御により実現することにより、実装を気にすることなく転送速度が向上する。すなわち、1本当りの転送速度は従来のバス方式と大差無いが、スイッチを中心にした場合、複数本のバスが並列に動作することが可能となる。また、スイッチを切り替えることでルーティングを行いディスクアレイ制御を行うことが可能となる。

【0030】

【実施例】以下、図面を用いて本発明の実施例を説明する。

【0031】（実施例1）

【0032】図1は、本発明に係る第1の実施例のハードウェア構成を示す。1はCPU、2はディスクアレイサブシステム、3はCPU1とディスクアレイサブシステムを結ぶネットワークである。

【0033】ディスクアレイサブシステム2は、図1に示すように、ディスクアレイ制御装置（DAC）4とn台のアレイモジュール9（9－1，9－2，…，9－n）で構成されている。DAC4は、複数の階層制御ルータ5とパリティ生成回路（PG）6とキャッシュメモリ7とスイッチ8で構成されている。階層制御ルータ5は、マイクロプロセッサで構成されている。階層制御ルータ5とパリティ生成回路（PG）6とキャッシュメモリ7とアレイモジュール9は、スイッチ8の各ポートに接続されている。

【0034】図2は、DAC4内のスイッチ8の内部構成を示している。n個（nは任意）の上位ポート20には、階層制御ルータ5やパリティ生成回路（PG）6やキャッシュメモリ7が接続される。m個（mは任意）の下位ポート21には、アレイモジュール9が接続される。

【0035】このスイッチ8では、上位、下位のポートにかかわらず、任意のポートの接続が可能である。

【0036】図3は、アレイモジュール9の内部構成を示す。アレイモジュール9は、下位RAID制御部（以下、単に下位RAIDと呼ぶ）10とローカルキャッシュメモリ11とドライブとのSCSI（small computer system interface）インターフェース制御を行うSPC12とが内部バス12に接続されて構成されている。各SPC12には、SCSIバスにより複数のドライブ14が接続されている。これらのドライブ14に対しては、SPC12のSCSIインターフェース制御により、データの読み出しまたは書き込みが行われる。

【0037】下位RAID10は、アレイモジュール9

の内部でのディスクアレイ制御を行うための下位ＲＡＩＤ制御用マイクロプロセッサ（ＭＰ）１８と、このＭＰ１８が使用するＭＰ用メモリ１５と、ＭＰ１８上で動作するマイクロプログラムが格納されるブートＲＯＭ１６と、スイッチ８とのインターフェース制御を行うインターフェース制御回路１７とで構成され、これらはすべて内部バス１２に接続されている。インターフェース制御回路１７は、線１９により、スイッチ８の下位ポート２１に接続される。

【００３８】このアレイモジュール９は、これ単独でディスクアレイとして機能する。すなわち、スイッチ８の下位ポート２１を経由して転送されてきたデータは、アレイモジュール９内のＭＰ１８により予め指定されたＲＡＩＤのレベルに従って当該ドライブ１４に格納される。このアレイモジュール９内のＲＡＩＤの制御は、従来より一般に行われている制御方法と同じで構わない。

【００３９】図１のディスクアレイサブシステム２では、下位側として従来のディスクアレイであるアレイモジュール９内でディスクアレイのＲＡＩＤ制御を行い、さらに上位側としてＤＡＣ４によりアレイモジュール９間でディスクアレイのＲＡＩＤ制御を行うことで、２段階のＲＡＩＤ制御を行うようにしている。本実施例の特徴は、このような２段階のＲＡＩＤ制御を行う際に、階層制御ルータ５によりスイッチ８の切り換えを制御することによって、上位側のＤＡＣ４におけるＲＡＩＤ制御を実現したことにある。以下に、その方法について説明する。

【００４０】図４は、ＤＡＣ４内の階層制御ルータ５が使用するルーティングテーブル２２の一例を示す。ルーティングテーブル２２は、階層制御ルータ５ごとに設けられ、ＤＡＣ４内のキャッシュメモリ７に格納されている。ＤＡＣ４内の各階層制御ルータ５は、ＣＰＵ１から読み出しまたは書き込み要求が発行された場合、スイッチ８を介してキャッシュメモリ７内のルーティングテーブル２２へアクセスする。

【００４１】なお、キャッシュメモリ７のようなスイッチ８のポートに接続したメモリにルーティングテーブル２２を格納する代わりに、各階層制御ルータ５から直接アクセス可能なメモリを用意し、そこにルーティングテーブル２２を格納する方法もある。このようにした場合、各階層制御ルータ５から直接アクセス可能なルーティングテーブル２２用のメモリを用意しなければならないが、ＣＰＵ１から読み出しまたは書き込み要求が発行される毎にスイッチ８を介してキャッシュメモリ７内のルーティングテーブル２２へアクセスする必要は無くなるので、スイッチ８を使用しなくてもよくなり、スイッチ８が性能のネックになる場合は有効である。

【００４２】本実施例では、ＣＰＵ１からの読み出しまたは書き込み要求はシーケンスとして送られてくる。シーケンスには、当該シーケンスを識別するためにＩＤが付けられる。そして、ルーティングテーブル２２を参照してこのシーケンスＩＤ２３に対応するエントリを認識することにより、ＤＡＣ４内の階層制御ルータ５はＲＡＩＤ制御を行う。シーケンスＩＤは、図４に示すように、当該処理の番号（シーケンス番号）と、読み出しまたは書き込みされるデータのアドレスで構成される。上位装置であるＣＰＵ１からは、このシーケンスＩＤに読み出しまたは書き込み（Ｒ／Ｗ）コマンドが付けられて送られてくる。

【００４３】図４は、ＲＡＩＤのレベル１のミラーリング（２重化）のＲＡＩＤ制御を行う場合に使用されるルーティングテーブル２２を示す。従来技術の欄で説明した特開平６－２４２８８７号の技術では、ＣＰＵとディスクアレイとを接続するネットワークスイッチではＲＡＩＤ制御を行わないため、このネットワークスイッチにおけるルーティング制御においては、ＣＰＵから送出される一つのシーケンスに対し送信先のポートは１個であった。これに対し、本実施例では、ＤＡＣ４において階層制御ルータ５がスイッチ８を切り替えることでＲＡＩＤ制御を実現する。このため、一つのシーケンスＩＤ２３に対し、階層制御ルータ５がルーティングテーブル２２を用いて複数の送信先ポート（例えば、図４の例では２重化を行うため２つの送信先ポート１　２４と送信先ポート２　２５）に変換する。

【００４４】本実施例では、ディスクアレイサブシステム２の初期設定の段階で、ＤＡＣ４およびアレイモジュールで行われるＲＡＩＤのレベルが設定される。つまり、ＤＡＣ４においては、どのアレイモジュール９のグループにより、どの様なレベルのＲＡＩＤ制御が行われるのかを初期設定の段階で設定し、アレイモジュール９においても、どのドライブ１４のグループにより、どの様なレベルのＲＡＩＤ制御が行われるのかを初期設定の段階で設定しておく。このため、ディスクアレイサブシステム２の使用開始時において、ルーティングテーブル２２は、送信先ポート１　２４および送信先ポート２　２５のみ設定されており、シーケンスＩＤ２３には何も登録されていない。そこで、このディスクアレイサブシステム２の使用を開始し、データの新規書き込みを行う場合は、ルーティングテーブル２２のシーケンスＩＤ２３に書き込むデータのシーケンスＩＤを登録し、ルーティングテーブル２２上でこの登録を行ったシーケンスＩＤ２３に対応する送信先ポート１　２４および送信先ポート２　２５に接続されているアレイモジュール９にデータを転送する。また、データを削除する場合は、ルーティングテーブル２２上において削除するデータのシーケンスＩＤ２３を削除し、同様に、アレイモジュール９内のテーブル上の当該シーケンスＩＤを削除する。

【００４５】なお、通常に使用している際に、新たに新しいデータを書き込みたい場合は、上述したような新規にデータを書き込む場合と同様に、ルーティングテーブ

ル２２上においてシーケンスＩＤ２３が登録されていない所を探し、シーケンスＩＤ２３が登録されていなければ、そこに書き込むデータのシーケンスＩＤを登録し、対応する送信先ポート１　２４および送信先ポート２　２５に接続されているアレイモジュール９にデータを転送する。このようにしてＤＡＣ４からデータを受け取ったアレイモジュール９では、同様に初期設定の段階で設定しておいたＲＡＩＤのレベルに従い、ドライブ１４にデータを書き込む。

【００４６】図５に、ＤＡＣ４において階層制御ルータ５がＲＡＩＤのレベル１（ＲＡＩＤ１）の制御を行う際のフローチャートを示す。階層制御ルータ５は、マイクロプロセッサで構成されている。

【００４７】まず、ＣＰＵ１より読み出しまたは書き込み要求が発生し、シーケンスＩＤが付けられて、ディスクアレイサブシステム２のＤＡＣ４に発行されたとする（ステップ２６）。本実施例では、予めＤＡＣ４で行うＲＡＩＤ制御のレベル（本例ではＲＡＩＤレベル１）をユーザが指定しておいてある。そこで、ＣＰＵ１から発行された、読み出しまたは書き込み要求を受け付けた階層制御ルータ５は、予め指定されているＲＡＩＤのレベルを認識する（ステップ２７）。図５の例では、ＲＡＩＤ１の制御を行うように認識する。

【００４８】次に、読み出しまたは書き込み要求を受け付けた階層制御ルータ５は、キャッシュメモリ７に格納されているルーティングテーブル２２を検索する（ステップ２８）。図５の例では、ＲＡＩＤ１の制御を行うためルーティングテーブル２２は図４のようになっている。そして、ＲＡＩＤ１のミラーリングの制御をＤＡＣ４の階層制御ルータ５が行うため、ＣＰＵ１から発行された１個の読み出しまたは書き込み要求は、図４のルーティングテーブル２２により、送信先ポート１２４と送信先ポート２　２５の２個のポートに変換される（ステップ２９）。「変換」とは、具体的には、新規データ書き込みの場合は、図４のルーティングテーブル２２からシーケンスＩＤ２３の欄が空きのエントリを探し、そこに書き込み要求のシーケンスＩＤを登録し、対応する送信先ポート１　２４と送信先ポート２　２５を認識することである。また、データの読み出しあるいは更新の場合は、読み出しまたは書き込み（更新）要求のシーケンスＩＤを図４のルーティングテーブル２２から探し、対応する送信先ポート１　２４と送信先ポート２　２５を認識することである。

【００４９】階層制御ルータ５が複数（２個）の送信先の下位ポート２１を認識したら、当該階層制御ルータ５が接続されている上位ポート２０と当該下位ポート２１を接続する（ステップ３０）。

【００５０】このとき、当該アレイモジュール９が他の読み出しまたは書き込み要求で使用中の場合は、使用可能になるまで待つ（ステップ３１）。当該アレイモジュール９が使用可能になったら、これらの当該アレイモジュール９が接続されている下位ポートに対し、読み出しまたは書き込み要求を発行する（ステップ３２）。つまり、階層制御ルータ５は、自分が接続されているスイッチ８の上位ポート２０と当該下位ポート２１（本例では２個の下位ポート）とを接続し、線１９を介して当該アレイモジュール９に対し読み出しまたは書き込み要求を発行することになる。

【００５１】このようにして読み出しまたは書き込み要求を受け取った当該アレイモジュール９では、アレイモジュール９内のＭＰ１８が、この要求の受付が可能かどうかを判断し、先に述べたようにアレイモジュール９内で独自に読み出しまたは書き込み処理を行い、ＣＰＵ１とアレイモジュール９間でデータの読み出しまたは書き込みが可能であれば、線１９、当該下位ポート２１、当該上位ポート２０を介し階層制御ルータ５に対し転送許可を発行する（ステップ３３）。この当該アレイモジュール９からの転送許可を受け取った階層制御ルータ５は、ＣＰＵ１と当該アレイモジュール９との間でデータ転送の制御を行う（ステップ３４）。

【００５２】上述の手順はデータの２重化を行うＲＡＩＤ１の制御を行うものであるため、同じデータが別のアレイモジュール９内に２重化されて格納される。このため、読み出し時は、当該データが格納されている２個のアレイモジュール９のうちで早く処理可能な方から当該データを読み出す。したがって、詳しく言えば、上記ステップ３０では２個の下位ポートのうちのどちらか一方と上位ポートを接続すればよい。一方、書き込み時は、両方のアレイモジュール９内に同じデータを転送し書き込む。

【００５３】（実施例２）

【００５４】実施例１では２重化を行うＲＡＩＤ１の制御を例に説明した。実施例２では、本発明をＲＡＩＤのレベル３（ＲＡＩＤ３）に適用した場合を説明する。従来技術の欄で説明したが、ＲＡＩＤ３では一つのデータを複数に分割しこれらを複数の別のドライブに並列に書き込む。このとき、ドライブの障害に対する信頼性を向上させるため、分割したデータからパリティを作成し、データとは別のドライブに格納する。読み出す場合は、逆に複数のドライブから分割されたデータを並列に読み出し、結合して上位のＣＰＵ１へ転送する。

【００５５】本実施例では、ＤＡＣ４の階層制御ルータ５が複数のアレイモジュール９に対し、ＤＡＣ４内のスイッチ８の切り替えを制御することにより、これらの複数のアレイモジュール９に対するＲＡＩＤ３の制御を行う。そこで、以下にその方法を説明する。なお、全体のハードウェア構成、スイッチの内部構成、およびアレイモジュールの内部構成については、上述の実施例１の図１、図２、および図３と同じであるので説明は省略する。

【００５６】図６は、ＲＡＩＤ３のときのＤＡＣ４内の階層制御ルータ５が使用するルーティングテーブル２２を示している。各階層制御ルータ５のルーティングテーブル２２の使用方法は、上述した実施例１のＲＡＩＤ１の場合と同じである。

【００５７】ＲＡＩＤ３においても、ルーティングテーブル２２におけるシーケンスＩＤ２３に対応するエントリを認識することにより、ＤＡＣ４内の階層制御ルータ５はＲＡＩＤ制御を行う。本実施例では、ＤＡＣ４において階層制御ルータ５がスイッチ８を切り替えることでＲＡＩＤ制御を実現する。このため、一つのシーケンスＩＤ２３に対し、階層制御ルータ５がルーティングテーブル２２を用いて複数の送信先ポート（図６の例ではＲＡＩＤ３制御を行うためデータ用の４つの送信先ポート１～４とパリティ用の送信先ポート５）に変換する。そして、送信先ポート１から４までに接続されているアレイモジュール９には分割されたデータが転送され、送信先ポート５に接続されているアレイモジュール９には分割されたデータから作成されたパリティが転送される。なお、パリティを作成せず、送信先ポート１から５に接続されている全てのアレイモジュール９に分割されたデータを転送することも可能である。

【００５８】図７に、ＤＡＣ４において階層制御ルータ５がＲＡＩＤ３の制御を行う際のフローチャートを示す。図７（ａ）は正常時の書き込み処理フローチャートを示し、図７（ｂ）は正常時の読み出し処理フローチャートを示し、図７（ｃ）は障害時の読み出し処理フローチャートを示している。

【００５９】まず、図７（ａ）により正常時の書き込み処理を以下に説明する。ＣＰＵ１より書き込み要求が発生し、シーケンスＩＤが付けられて、ディスクアレイサブシステム２のＤＡＣ４に発行されたとする（ステップ４０）。本実施例では、予めＤＡＣ４で行うＲＡＩＤ制御のレベル（本例ではＲＡＩＤ３）をユーザが指定しておいてある。そこで、ＣＰＵ１から発行された、書き込み要求を受け付けた階層制御ルータ５は、予め指定されているＲＡＩＤのレベルを認識する（ステップ４１）。図７の例ではＲＡＩＤ３の制御を行うように認識する。

【００６０】次に、書き込み要求を受け付けた階層制御ルータ５は、キャッシュメモリ７に格納されているルーティングテーブル２２を検索する（ステップ４２）。本実施例では、ＲＡＩＤ３の制御を行うためルーティングテーブル２２は図６のようになっている。そして、ＲＡＩＤ３の制御をＤＡＣ４の階層制御ルータ５が行うため、ＣＰＵ１から発行された１個の書き込み要求は、図６のルーティングテーブル２２により、分割されたデータが書き込まれるアレイモジュール９が接続された送信先ポート１～４と、これらの分割されたデータから作成されたパリティが書き込まれるアレイモジュール９が接続された送信先ポート５との５個のポートに変換され、同時にＰＧ６が接続されているポートを認識する（ステップ４３）。

【００６１】次に、ＰＧ６の使用状況を調べ、使用可能な場合は、当該階層制御ルータ５が接続されている上位ポート２０とＰＧ６が接続されている上位ポート（ＰＧポート）とを接続する（ステップ４４）。当該階層制御ルータ５が接続されているスイッチ８の上位ポート２０とＰＧポートとの接続が完了したら、ＣＰＵ１から当該階層制御ルータ５およびＰＧポートを介してＰＧ６へデータを転送し、ＰＧ６においてデータの分割を行い、この分割したデータからパリティを生成する（ステップ４５）。次に、当該階層制御ルータ５は、ＰＧポートと当該下位ポート２１（データ用の送信先ポート１～４とパリティ用の送信先ポート５）を接続するようにＰＧポートに指示する（ステップ４６）。

【００６２】このとき、当該アレイモジュール９が、他の読み出しまたは書き込み要求で使用中の場合は使用可能になるまで待つ（ステップ４７）。当該アレイモジュール９が使用可能になったら、これらの当該アレイモジュール９が接続されている５個の下位ポートに対し、書き込み要求を発行する（ステップ４８）。つまり、階層制御ルータ５は、自分が接続されている上位ポートとＰＧポートを接続し、ＰＧポート経由でこれらの当該アレイモジュール９が接続されている下位ポートに対し、書き込み要求を発行することになる。

【００６３】このようにして書き込み要求を受け取った当該アレイモジュール９では、アレイモジュール９内のＭＰ１８が、この要求の受付が可能かどうかを判断し、書き込みが可能な場合は、線１９、当該下位ポート２１、ＰＧポート、当該上位ポート２０を介し階層制御ルータ５に対し転送許可を発行する（ステップ４９）。この当該アレイモジュール９からの転送許可を階層制御ルータ５が受け取ったら、当該階層制御ルータ５の制御の元で、ＰＧ６から当該アレイモジュール９へ、分割されたデータおよびこれらのデータから作成されたパリティを転送し、アレイモジュール９内では独自に書き込み処理を行う（ステップ５０）。

【００６４】次に、図７（ｂ）により正常時の読み出し処理を以下に説明する。ＣＰＵ１より読み出し要求が発生し、シーケンスＩＤが付けられて、ディスクアレイサブシステム２のＤＡＣ４に発行されたとする（ステップ５１）。ステップ５２、５３、５４は、上述した正常時の書き込み処理のステップ４１、４２、４３とそれぞれ同じである。

【００６５】階層制御ルータ５が複数の送信先下位ポート２１とＰＧポートを認識したら、当該階層制御ルータ５は、ＰＧポートに対し、当該下位ポート２１との接続を指示する（ステップ５５）。

【００６６】当該アレイモジュール９が、他の読み出しまたは書き込み要求で使用中の場合は、使用可能になる

まで待つ（ステップ５６）。当該アレイモジュール９が使用可能になったら、当該階層制御ルータ５は、自分が接続されている上位ポートとＰＧポートとを接続し、ＰＧポートを経由してこれらの当該アレイモジュール９が接続されている下位ポートに対し、読み出し要求を発行する（ステップ５７）。

【００６７】このようにして読み出し要求を受け取った当該アレイモジュール９では、アレイモジュール９内のＭＰ１８が、この要求の受付が可能かを判断し、アレイモジュール９内では独自に読み出し処理を行い、ＣＰＵ１とアレイモジュール９間でデータの読み出しが可能な場合は、線１９、当該下位ポート２１、ＰＧポート、および当該上位ポート２０を介して階層制御ルータ５に対し転送許可を発行する（ステップ５８）。この当該アレイモジュール９からの転送許可を受け取った階層制御ルータ５は、当該アレイモジュール９から送出される分割されたデータをＰＧ６へ転送するように制御する（ステップ５９）。このようにして各当該アレイモジュール９から受け取った分割されたデータはＰＧ６において結合され、結合されたデータはＰＧ６からＰＧポートおよび階層制御ルータ５を経由してＣＰＵ１へ転送される（ステップ６０）。

【００６８】次に、図７（ｃ）により障害時の読み出し処理を以下に説明する。まず、本実施例における障害について説明する。本実施例では、アレイモジュール９内のドライブ１３の障害はアレイモジュール９内で対策されているものとする。つまり、あるアレイモジュール９内のドライブ１３に障害が発生しており、この障害が発生しているドライブに読み出し要求が発行された場合は、アレイモジュール９内の残りの正常なドライブ１３に格納されているデータとパリティから、障害ドライブ内のデータを回復して、あたかも正常であるかのようにデータを転送してくる。このように、本実施例のＤＡＣ４では、アレイモジュール９内のドライブ障害に対してはアレイモジュール９に任せ、対応しない。

【００６９】以下の図７（ｃ）で説明するＤＡＣ４の階層制御ルータ５が対応する障害は、アレイモジュール９自身では対応できないような障害である。例えば、アレイモジュール９の制御部に障害が発生し、アレイモジュール９内の全てのドライブ１３に対し読み出しまたは書き込みができないような、アレイモジュール９全体に及ぶような障害とする。

【００７０】そこで、以下に具体的な処理方法を説明する。

【００７１】ＣＰＵ１より読み出し要求が発生し、シーケンスＩＤが付けられて、ディスクアレイサブシステム２のＤＡＣ４に発行されたとする（ステップ６１）。ステップ６２、６３は、上述した図７（ｂ）の正常時の読み出し処理のステップ５２、５３とそれぞれ同じである。本実施例では、ＲＡＩＤ３の制御を行うためルーティングテーブル２２は図６のようになっている。そして、ＲＡＩＤ３の制御をＤＡＣ４の階層制御ルータ５が行うため、ＣＰＵ１から発行された１個の読み出し要求は、図６のルーティングテーブル２２により、分割されたデータが書き込まれるアレイモジュール９が接続された送信先ポート１～４と、これらの分割されたデータから作成されたパリティが書き込まれるアレイモジュール９が接続された送信先ポート５との５個のポートに変換され、同時にＰＧ６が接続されているポートを認識する（ステップ６４）。

【００７２】なお、アレイモジュール９に障害が発生した場合、ルーティングテーブル２２では、障害が発生したアレイモジュール９が接続されている下位ポート名にフラグが付き、階層制御ルータ５はこれによりアレイモジュール９の障害の有無を判断することが可能である。

【００７３】次に、ＰＧ６の使用状況を調べ、使用可能な場合は、当該アレイモジュール９（障害が発生したアレイモジュール９は除く）が接続されている下位ポート２１とＰＧポートとを接続する（ステップ６５）。このとき、当該アレイモジュール９が、他の読み出しまたは書き込み要求で使用中の場合は使用可能になるまで待つ（ステップ６６）。当該アレイモジュール９が使用可能になったら、当該階層制御ルータ５は、自分が接続されている上位ポートとＰＧポートとを接続し、ＰＧポートを経由してこれらの当該アレイモジュール９が接続されている下位ポートに対し、読み出し要求を発行する（ステップ６７）。

【００７４】このようにして読み出し要求を受け取った当該アレイモジュール９では、アレイモジュール９内のＭＰ１８が、この要求の受付が可能かどうかを判断し、アレイモジュール９内では独自に読み出し処理を行い、アレイモジュール９でデータの読み出しが可能な場合は、線１９、当該下位ポート２１、ＰＧポート、および上位ポート２０を介して階層制御ルータ５に対し転送許可を発行する（ステップ６８）。この当該アレイモジュール９からの転送許可を受け取った階層制御ルータ５は、当該アレイモジュール９から送出される分割されたデータおよびパリティをＰＧ６に転送するように制御する（ステップ６９）。

【００７５】ＰＧ６では、当該アレイモジュール９から当該下位ポート２１およびＰＧポートを介してＰＧ６へ転送されたデータおよびパリティを用いて、障害が発生したアレイモジュール９内に格納されているデータを回復し、この回復したデータとアレイモジュール９から転送されてきた分割されたデータとを結合する（ステップ７０）。このようにして結合されたデータは、ＰＧポートと階層制御ルータ５が接続されている上位ポート２０を結合し、階層制御ルータ５を介してＣＰＵ１へ転送される（ステップ７１）。

【００７６】また、もし予備のアレイモジュール９があ

る場合、または障害が発生したアレイモジュール９を正常なアレイモジュール９に交換した場合は、ＰＧ６で回復した障害が発生したアレイモジュール９内のデータを、ＣＰＵ１ではなく予備のアレイモジュール９または交換した正常なアレイモジュール９に転送し、障害が発生したアレイモジュール９の復元を行うことも可能である。この制御は、所定の階層制御ルータ５が行うようにすればよい。

【００７７】（実施例３）

【００７８】次に、本発明をＲＡＩＤのレベル５（ＲＡＩＤ５）に適用した場合を以下に示す。従来技術の欄で説明したが、ＲＡＩＤ５では一つのデータを分割せずに１台のドライブに格納する。このとき、ドライブの障害に対する信頼性を向上させるため、各ドライブのデータからパリティを作成し、データとは別のドライブに格納する。このとき、パリティを格納するドライブを特定の１台に固定した場合はＲＡＩＤ４になり、特定せずにデータと同様に複数のドライブに分散させた場合はＲＡＩＤ５になる。ＲＡＩＤ５の場合は、データは分割されていないため、ディスクアレイを構成するすべてのドライブが独立に動作することが可能となり、単位時間当りに処理することが可能な読み出し処理件数は向上する。しかし、書き込み時には、パリティを更新するために、書き込まれるアドレスにすでに書き込まれているデータ（旧データ）とパリティ（旧パリティ）を読み出し、新しいパリティを作成した後にデータおよびパリティを書き込む必要がある。したがって、１回の書き込み処理に対して、２回の読み出しと２回の書き込み処理が必要となり、このオーバヘッドが問題となっている。

【００７９】ＲＡＩＤ５では、先に述べたように書き込み時にはパリティを更新するため、旧データ、旧パリティの２回の読み出しが必要になるため、ＲＡＩＤ５の制御を行う場合は、ＣＰＵ１から書き込み要求が発行されたら階層制御ルータ５は独自に旧データ、旧パリティの２回の読み出し要求を発行する。そこで、以下にその方法を説明する。なお、全体のハードウェア構成、スイッチの内部構成、およびアレイモジュールの内部構成については、上述の実施例１の図１、図２、および図３と同じであるので説明は省略する。

【００８０】図８は、ＲＡＩＤ５のときのＤＡＣ４内の階層制御ルータ５が使用するルーティングテーブル２２を示している。各階層制御ルータ５のルーティングテーブル２２の使用方法は、上述した実施例１、２と同じである。

【００８１】ＲＡＩＤ５においても、ルーティングテーブル２２におけるシーケンスＩＤ２３に対応するエントリを認識することにより、ＤＡＣ４内の階層制御ルータ５はＲＡＩＤ制御を行う。本実施例では、ＤＡＣ４において階層制御ルータ５がスイッチ８を切り替えることでＲＡＩＤ制御を実現する。このため、一つのシーケンスＩＤ２３に対し、階層制御ルータ５がルーティングテーブル２２を用いて複数の送信先ポート（図８の例ではＲＡＩＤ５制御を行うためデータの送信先ポートとパリティの送信先ポートの２個のポート）に変換する。

【００８２】図９および図１０に、ＤＡＣ４において階層制御ルータ５がＲＡＩＤ５の制御を行う際のフローチャートを示す。図９は正常時の書き込み処理フローチャートを示し、図１０（ａ）は正常時の読み出し処理フローチャートを示し、図１０（ｂ）は障害時の読み出し処理フローチャートを示している。

【００８３】まず、図９により正常時の書き込み処理を以下に説明する。ＣＰＵ１より書き込み要求が発生し、シーケンスＩＤが付けられて、ディスクアレイサブシステム２のＤＡＣ４に発行されたとする（ステップ７５）。本実施例では、予めＤＡＣ４で行うＲＡＩＤ制御のレベル（本例ではＲＡＩＤ５）をユーザが指定しておいてある。そこで、ＣＰＵ１から発行された、書き込み要求を受け付けた階層制御ルータ５は、予め指定されているＲＡＩＤのレベルを認識する（ステップ７６）。図９の例ではＲＡＩＤ５の制御を行うように認識する。

【００８４】次に、書き込み要求を受け付けた階層制御ルータ５は、キャッシュメモリ７に格納されているルーティングテーブル２２を検索する（ステップ７７）。本実施例ではＲＡＩＤ５の制御を行うためルーティングテーブル２２は図８のようになっている。そして、ＲＡＩＤ５の制御をＤＡＣ４の階層制御ルータ５が行うため、ＣＰＵ１から発行された１個の書き込み要求は、図８のルーティングテーブル２２により、データが書き込まれるアレイモジュール９が接続された送信先ポート１　７３とパリティが書き込まれるアレイモジュール９が接続された送信先ポート２　７４との２個のポートに変換され、同時にＰＧ６が接続されているＰＧポートを認識する（ステップ７８）。

【００８５】なお、図８のルーティングテーブル２２において、送信先ポート１　７３がデータの送信先ポートであり、送信先ポート２　７４がパリティの送信先ポートである。また、送信先ポート３と送信先ポート４は、送信先ポート２　７４に接続されたアレイモジュール９に格納されているパリティを作成するのに使用したデータが格納されているアレイモジュール９が接続されているポートを示す。つまり、送信先ポート１、３、４に接続されているアレイモジュール９内のデータから作成されたパリティが、送信先ポート２に接続されているアレイモジュール９内に格納されている。したがって、正常時の書き込み処理では、送信先ポート１のアレイモジュール９から旧データを読み出し、送信先ポート２のアレイモジュール９から旧パリティを読み出し、読み出した旧データおよび旧パリティと新規書き込みデータとを用いて新パリティを作成し、新規書き込みデータを送信先ポート１のアレイモジュール９に書き込み、新パリティ

を送信先ポート２のアレイモジュール９に書き込むことになる。送信先ポート３、４は後述する障害時に使用する。

【００８６】ステップ７８の後、階層制御ルータ５は、自分が接続されている上位ポート２０とＰＧポートとをスイッチ８を切り替えて接続する（ステップ７９）。次に、ＣＰＵ１からの書き込みデータを、上位ポート２０とＰＧポートを介してＰＧ６に転送する（ステップ８０）。

【００８７】ＲＡＩＤ５では書き込み時においてパリティを更新するため、旧データ、旧パリティの２回の読み出しが必要になる。このため、ＲＡＩＤ５の場合、階層制御ルータ５は、独自に旧データ、旧パリティの読み出し要求を発行する必要がある。以下のステップ８１～８３は、そのための処理である。

【００８８】ステップ７８で階層制御ルータ５は既にデータおよびパリティの送信先ポートである下位ポート２１を認識しているため、当該階層制御ルータ５は、ＰＧポートと当該下位ポート２１（旧データ用の送信先ポート１と旧パリティ用の送信先ポート２）とを接続する（ステップ８１）。

【００８９】このとき、当該アレイモジュール９が、他の読み出しまたは書き込み要求で使用中の場合は使用可能になるまで待つ（ステップ８２）。当該アレイモジュール９が使用可能になったら、旧データおよび旧パリティが格納されているこれらの当該アレイモジュール９が接続されている下位ポートに対し、読み出し要求を発行する（ステップ８３）。

【００９０】このようにして読み出し要求を受け取った各当該アレイモジュール９では、アレイモジュール９内のＭＰ１８が、この要求の受付が可能かを判断し、ＣＰＵ１とアレイモジュール９間でデータの読み出しが可能な場合は、線１９、当該下位ポート２１、ＰＧポート、および当該上位ポート２０を介して階層制御ルータ５に対し転送許可を発行する（ステップ８４）。この当該アレイモジュール９からの転送許可を受け取った階層制御ルータ５は、当該アレイモジュール９から旧データおよび旧パリティを受け取る（ステップ８５）。ＰＧ６では、このようにして各当該アレイモジュール９から受け取った旧データと旧パリティ、およびステップ８０において既にＣＰＵ１からＰＧ６に転送されている新データを用いて、新パリティを生成する（ステップ８６）。

【００９１】次に、当該アレイモジュール９が、他の読み出しまたは書き込み要求で使用中の場合は使用可能になるまで待つ（ステップ８８）。当該アレイモジュール９が使用可能になったら、階層制御ルータ５は、これらの当該アレイモジュール９が接続されている下位ポート（送信先ポート１、２）に対し、新データおよび新パリティの書き込み要求を発行する（ステップ８９）。

【００９２】このようにして書き込み要求を受け取った当該アレイモジュール９では、アレイモジュール９内のＭＰ１８が、この要求の受付が可能かを判断し、書き込みが可能な場合は、線１９、当該下位ポート２１、ＰＧポート、および当該上位ポート２０を介して階層制御ルータ５に対し転送許可を発行する（ステップ９０）。この当該アレイモジュール９からの転送許可を階層制御ルータ５が受け取ったら、当該階層制御ルータ５の制御の元で、ＰＧ６から当該アレイモジュール９へ、新データおよび作成された新パリティを転送し、アレイモジュール９内では独自に書き込み処理を行う（ステップ９１）。

【００９３】次に、図１０（ａ）により正常時の読み出し処理を以下に説明する。ＣＰＵ１より読み出し要求が発生し、シーケンスＩＤが付けられて、ディスクアレイサブシステム２のＤＡＣ４に発行されたとする（ステップ９２）。ステップ９３、９４は、上述した正常時の書き込み処理のステップ７６、７７とそれぞれ同じである。

【００９４】正常時の読み出し処理では、ＣＰＵ１から発行された１個の読み出し要求は、図８のルーティングテーブル２２により、データが書き込まれているアレイモジュール９が接続された下位ポート２１（送信先ポート１）に変換される（ステップ９５）。ステップ９５で階層制御ルータ５が当該データが書き込まれているアレイモジュール９が接続された下位ポート２１を認識したら、当該階層制御ルータ５が接続されている上位ポート２０と当該下位ポート２１とを接続する（ステップ９６）。

【００９５】このとき、当該アレイモジュール９が、他の読み出しまたは書き込み要求で使用中の場合は、使用可能になるまで待つ（ステップ９７）。当該アレイモジュール９が使用可能になったら、当該アレイモジュール９が接続されている下位ポートに対し、読み出し要求を発行する（ステップ９８）。

【００９６】このようにして読み出し要求を受け取った当該アレイモジュール９では、アレイモジュール９内のＭＰ１８が、この要求の受付が可能かを判断し、ＣＰＵ１とアレイモジュール９間でデータの読み出しが可能な場合は、線１９、当該下位ポート２１、および当該上位ポート２０を介して階層制御ルータ５に対し転送許可を発行する（ステップ９９）。この当該アレイモジュール９からの転送許可を受け取った階層制御ルータ５は、当該アレイモジュール９から当該データを受け取りＣＰＵ１へ転送する（ステップ１００）。

【００９７】次に、図１０（ｂ）により障害時の読み出し処理を以下に説明する。ＲＡＩＤ５の障害も、ＲＡＩＤ３と同様に、ＤＡＣ４の階層制御ルータ５が対応する障害はアレイモジュール９自身では対応できないような障害である。そこで、以下に具体的な処理方法を説明する。

【００９８】ＣＰＵ１より障害が発生したアレイモジュール９内に書き込まれているデータに読み出し要求が発生し、シーケンスＩＤが付けられて、ディスクアレイサブシステム２のＤＡＣ４に発行されたとする（ステップ１０１）。ステップ１０２、１０３は、上述した正常時の読み出し処理のステップ９３、９４とそれぞれ同じである。

【００９９】ＲＡＩＤ５における障害時の読み出し処理では、当該データが作成に関与したパリティと、このパリティ作成に関与した当該データ以外の全データを読み出して、これらから当該データの回復を行う。このため、これらのデータとパリティが書き込まれているアレイモジュール９が接続されている全下位ポート２１とＰＧ６が接続されているＰＧポートを認識する。

【０１００】具体的には、ルーティングテーブル２２は図８のようになっており、読み出し要求のシーケンスＩＤに対応する送信先ポート１の下位ポートに接続されているアレイモジュール９に障害が発生している。そこで、当該シーケンスＩＤに対応する送信先ポート２～４を認識することになる。なお、図８のルーティングテーブル２２において、例えば、シーケンスＩＤ１で書き込み要求されたデータは下位ポート１に接続されたアレイモジュール９に書き込まれ、シーケンスＩＤ２で書き込み要求されたデータは下位ポート３に接続されたアレイモジュール９に書き込まれ、シーケンスＩＤ３で書き込み要求されたデータは下位ポート４に接続されたアレイモジュール９に書き込まれ、さらにこれらの書き込みデータから作成されたパリティが下位ポート２に書込まれている。したがって、正常時にはシーケンスＩＤ２、３で下位ポート３、４（送信先ポート１）から読み出していたデータを、障害時にはシーケンスＩＤ１で下位ポート３、４（送信先ポート３、４）から読み出すことになる。そのため、本実施例では、これらのパリティの作成に用いた複数のデータのシーケンスＩＤ（図４）は、異なるシーケンス番号で同じデータアドレスからなるシーケンスＩＤとし、各アレイモジュール９内ではデータアドレスのみでデータを特定できるようにしてある。したがって、上記シーケンスＩＤ１～３の例では、これらのシーケンスＩＤ１～３を異なるシーケンス番号で同じデータアドレスから構成されるようにし、正常時にはシーケンスＩＤ２、３で下位ポート３、４（送信先ポート１）から読み出していたデータを、障害時にはシーケンスＩＤ１で下位ポート３、４（送信先ポート３、４）から読み出すことができるようにしてある。なお、別の方法を用いることもできる。例えば、図８の送信先ポート３、４の欄に当該下位ポートに書き込んだデータのシーケンスＩＤを書く欄を加えておくようにしてもよい。

【０１０１】ステップ１０３の後、ＣＰＵ１から発行された１個の読み出し要求は、図８のルーティングテーブル２２により、パリティの作成に関与したデータとそのパリティが書き込まれているアレイモジュール９が接続されている送信先ポート２～４に変換され、同時にＰＧ６が接続されているポートを認識する（ステップ１０４）。

【０１０２】次に、ＰＧ６の使用状況を調べ、使用可能な場合は、当該アレイモジュール９が接続されている下位ポート２１とＰＧポートとを接続する（ステップ１０５）。このとき、当該アレイモジュール９が、他の読み出しまたは書き込み要求で使用中の場合は使用可能になるまで待つ（ステップ１０６）。当該アレイモジュール９が使用可能になったら、これらの当該アレイモジュール９が接続されている下位ポートに対し、読み出し要求を発行する（ステップ１０７）。

【０１０３】このようにして読み出し要求を受け取った当該アレイモジュール９では、アレイモジュール９内のＭＰ１８が、この要求の受付が可能かどうかを判断し、アレイモジュール９でデータの読み出しが可能な場合は、線１９、当該下位ポート２１、ＰＧポート、および上位ポート２０を介して階層制御ルータ５に対し転送許可を発行する（ステップ１０８）。この当該アレイモジュール９からの転送許可を受け取った階層制御ルータ５は、当該アレイモジュール９から送出されるデータおよびパリティをＰＧ６に転送するように制御する（ステップ１０９）。当該アレイモジュール９より当該下位ポート２１およびＰＧポートを介してＰＧ６へ転送されたデータおよびパリティにより、ＰＧ６は、障害が発生したアレイモジュール９内に格納されているデータを回復し、この回復したデータは、ＰＧポートと階層制御ルータ５が接続されている上位ポート２０を結合し、階層制御ルータ５を介してＣＰＵ１へ転送される（ステップ１１０）。

【０１０４】また、もし予備のアレイモジュール９がある場合、または障害が発生したアレイモジュール９を正常なアレイモジュール９に交換した場合は、ＰＧ６で回復した障害が発生したアレイモジュール９内のデータを、ＣＰＵ１ではなく予備のアレイモジュール９または交換した正常なアレイモジュール９に転送し、障害が発生したアレイモジュール９の復元を行うことも可能である。この制御は、所定の階層制御ルータ５が行うようにすればよい。

【０１０５】（変形例）

【０１０６】図１１は、図１のハードウェア構成の変形例を示す構成図である。図１１では図１と異なり、ＣＰＵ１が階層制御ルータ５を介さずに直接ＤＡＣ４のスイッチ８の上位ポート２０に接続されている。このとき、ＣＰＵ１からディスクアレイサブシステム２へ読み出しまたは書き込み要求が発行された場合、読み出しまたは書き込み要求を発行したＣＰＵ１に接続されている上位ポート２０は、必ず階層制御ルータ５が接続されている上位ポート２０にスイッチ８を切り換えて接続し、ＣＰ

Ｕ１からの読み出しまたは書き込み要求は階層制御ルータ５で受け付けられる。階層制御ルータ５がＣＰＵ１からの読み出しまたは書き込み要求を受け付けた後は、図１と同様に実施例１、２、３で示したようにＲＡＩＤ制御が行われる。

【０１０７】以上の実施例では、図１に示すように、ＤＡＣ４とｎ台のアレイモジュール９が１つの筐体に内蔵されてディスクアレイサブシステム２内を構成しているが、図１２に示すように、ＤＡＣ４の筐体とｎ台のアレイモジュール９が格納されているアレイドライブユニット１１２の筐体とを分離することも可能である。これは、ＤＡＣ４をバスではなくスイッチ８を中心に構成し、このスイッチ８に階層制御ルータ５を介してＣＰＵ１やアレイモジュール９を線で接続するようにしたためである。これにより、ＤＡＣ４とアレイドライブユニット１１２とを別の場所に設置することが可能になる。

【０１０８】また、この拡張として、ｎ台のアレイモジュール９を１ケ所に集めてアレイドライブユニット１１２の筐体内に収めるのではなく、ｎ台のアレイモジュール９を各々独立に設置することも可能である。このようにすると、ｎ台のアレイモジュール９をそれぞれ別の場所に設置することが可能になる。

【０１０９】図１３は、複数のＤＡＣ４でアレイモジュール９を共有する形態を示す。本発明は、このようにアレイモジュール９を複数のＤＡＣ４が共有した構成においても適用可能である。実現の方式としては、例えば各アレイモジュールにおけるデータ格納領域をＤＡＣ＃１用の領域とＤＡＣ＃２用の領域との２つに分けておき、ＤＡＣ＃１からは各アレイモジュールのＤＡＣ＃１用の領域にアクセスし、ＤＡＣ＃２からは各アレイモジュールのＤＡＣ＃２用の領域にアクセスするようにする方式がある。また、データ格納領域を分けずに、ＤＡＣ＃１とＤＡＣ＃２とで同じルーティングテーブルを共有して（例えば、共通にアクセスできるメモリを設けて、そこにルーティングテーブルを共有する）制御するようにしてもよい。

【０１１０】このようにアレイモジュール９を複数のＤＡＣ４が共有することにより、ＤＡＣ４に障害が発生した場合や、ＣＰＵ１とＤＡＣ４との間の線１１３やＤＡＣ４とアレイモジュール９との間の線１１４に障害が発生した場合に、もう一方のＤＡＣ４やＣＰＵ１とＤＡＣ４との間の線１１３やＤＡＣ４とアレイモジュール９との間の線１１４を使用することで、アレイモジュール９への読み出しまたは書き込みが可能となり、信頼性や可用性を向上させることが可能となる。また、このようにすると、アレイモジュール９への読み出しまたは書き込み処理を行えるバスが増加するので、単位時間当りに処理可能な読み出しまたは書き込み要求数を増加させることが可能となる。つまり、このアレイモジュール９へのバスがネックになるような場合は、このような構成が有効となる。

【０１１１】なお、上記実施例ではキャッシュについて考慮することなく説明をしたが、図１のキャッシュメモリ７に読み出すべきデータがキャッシングされているときはキャッシュメモリ７からデータを読み出し、キャッシュメモリ７に読み出すべきデータが無い場合に上述の読み出し処理を行うようにしても良い。また、書き込みについても、一旦キャッシュメモリ７にデータを書き込み、後で上述の書き込み処理でキャッシュメモリ７からドライブへデータの書き込みを行うようにしても良い。

【０１１２】

【発明の効果】複数のディスクアレイモジュール間によるディスクアレイ制御をクロスバ方式などのスイッチ手段のルーティング制御により実現することにより、実装を気にすることなく転送速度が向上する。すなわち、１本当りの転送速度は従来のバス方式と大差無いが、スイッチを中心にした場合、複数本のバスが並列に動作することが可能となる。したがって、転送速度を向上させることができる。また、スイッチを切り替えることでルーティングを行いディスクアレイ制御を行うことで、バスのようにアービトレーションのような制御が不要になるため制御が簡単になる。

【図面の簡単な説明】

【図１】本発明の実施例のハードウェア構成を示す図である。

【図２】図１のＤＡＣ内のスイッチの内部構造を示した図である。

【図３】図１のアレイモジュールの内部構造を示した図である。

【図４】実施例１のルーティングテーブルを説明する図である。

【図５】実施例１の読み出しおよび書き込み処理のタイミングチャートを示す図である。

【図６】実施例２のルーティングテーブルを説明する図である。

【図７】実施例２の読み出しおよび書き込み処理のタイミングチャートを示す図である。

【図８】実施例３のルーティングテーブルを説明する図である。

【図９】実施例３の正常時の書き込み処理のタイミングチャートを示す図である。

【図１０】実施例３の正常時および障害時の読み出し処理のタイミングチャートを示す図である。

【図１１】実施例の変形例のハードウェア構成を示す図である。

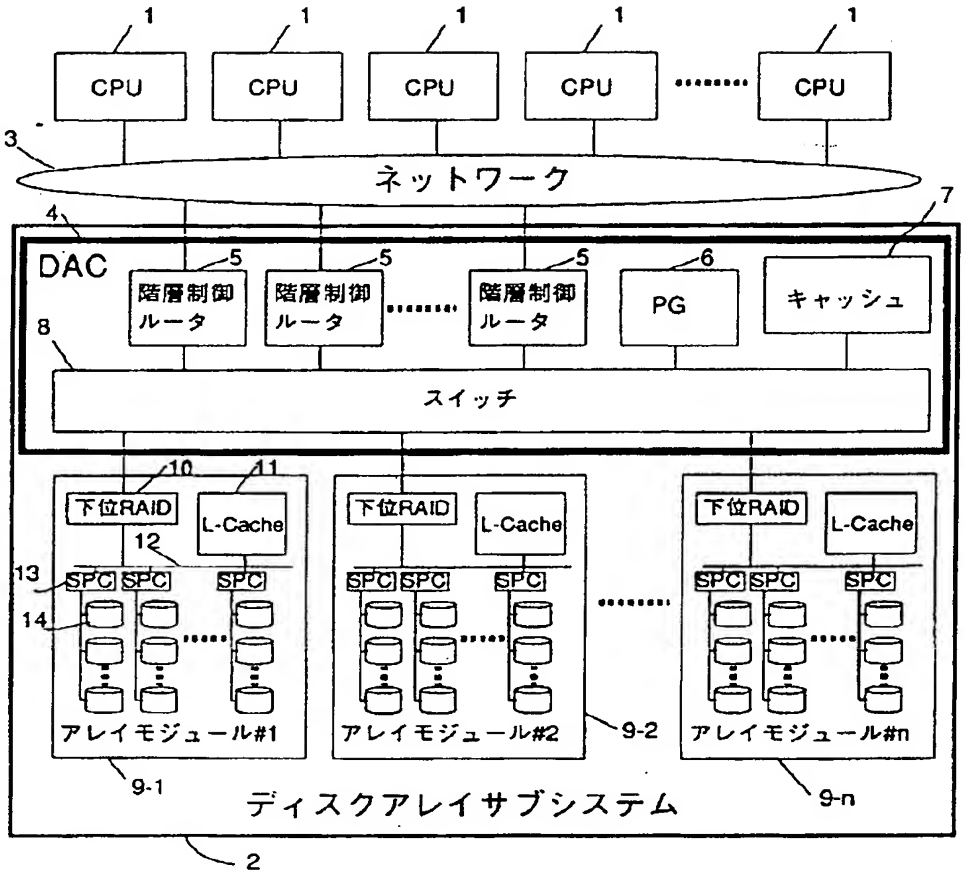
【図１２】実施例の変形例のハードウェア構成を示す図である。

【図１３】実施例の変形例のハードウェア構成を示す図である。

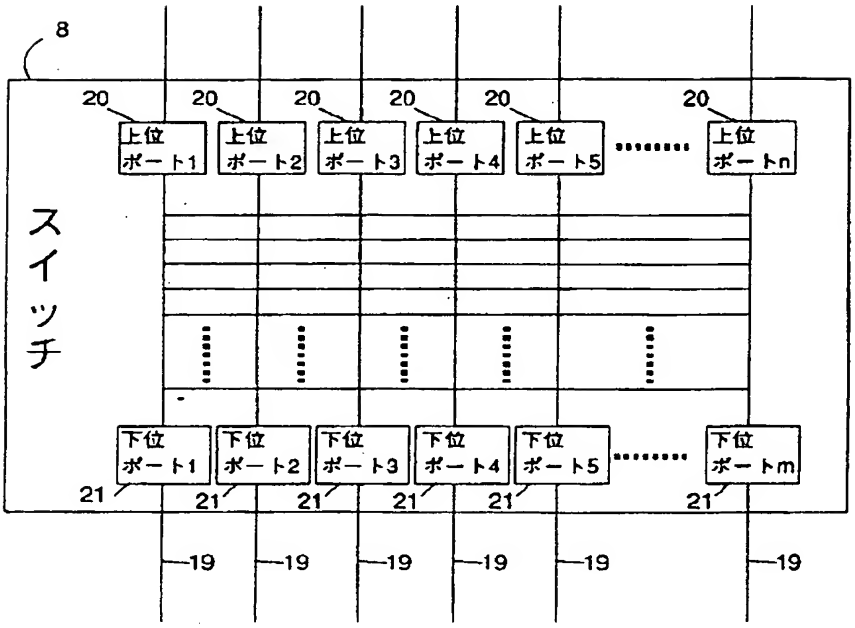
【符号の説明】

１…ＣＰＵ、２…ディスクアレイサブシステム、３…ネットワーク、４…ディスクアレイコントローラ（ＤＡＣ）、５…階層制御ルータ、６…パリティ生成回路（ＰＧ）、７…キャッシュメモリ、８…スイッチ、９…アレイモジュール、１０…下位ＲＡＩＤコントローラ、１２…内部バス、１３…ＳＰＣ、１４…ドライブ、１５…ＭＰ用メモリ、１６…Ｂｏｏｔ　ＲＯＭ、１７…インターフェース制御回路、１８…下位ＲＡＩＤ制御用ＭＰ（Ｍ＊Ｐ）、１９…線、２０…上位ポート、２１…下位ポート、２２…ＲＡＩＤ１用ルーティングテーブル、２３…シーケンスＩＤ、２４…送信先ポート１、２５…送信先ポート２、３５…ＲＡＩＤ３用ルーティングテーブル、７２…ＲＡＩＤ５用ルーティングテーブル、１１２…アレイドライブユニット、１１３…ＣＰＵとＤＡＣ間の線、１１４…ＤＡＣとアレイモジュール間の線。

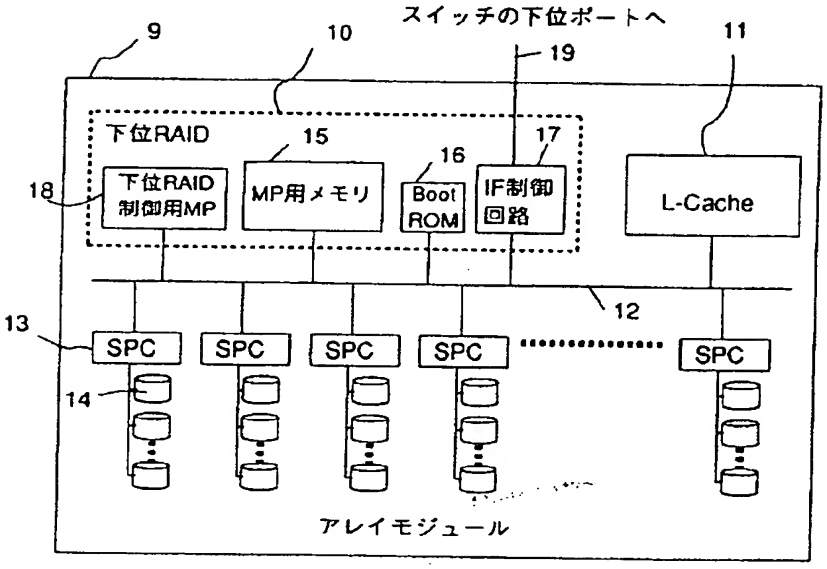
【図１】



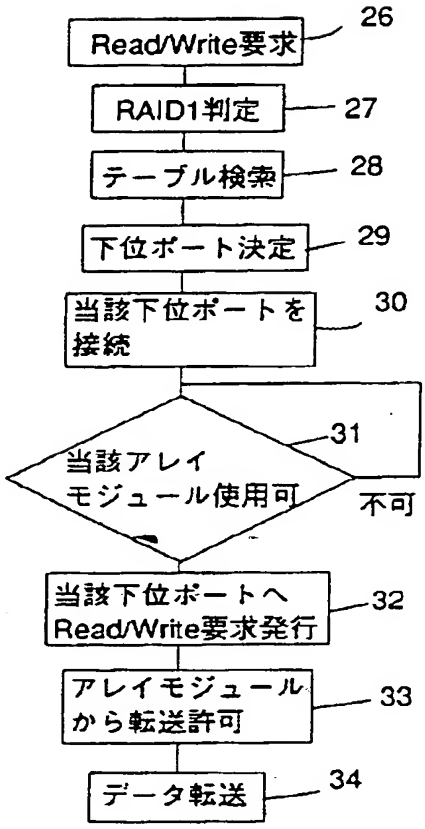
【図２】


8
スイッチ
20
上位ポート1
上位ポート2
上位ポート3
上位ポート4
上位ポート5
上位ポートn
下位ポート1
下位ポート2
下位ポート3
下位ポート4
下位ポート5
下位ポートm
21
19

【図３】


スイッチの下位ポートへ
9
10
11
19
下位RAID
15
16
17
18
下位RAID制御用MP
MP用メモリ
Boot ROM
IF制御回路
L-Cache
12
13
SPC
14
アレイモジュール

【図５】


Read/Write要求
26
RAID1判定
27
テーブル検索
28
下位ポート決定
29
当該下位ポートを接続
30
当該アレイモジュール使用可
31
不可
当該下位ポートへRead/Write要求発行
32
アレイモジュールから転送許可
33
データ転送
34

【図４】

シーケンスID

シーケンス番号	データアドレス	R/Wコマンド

RAID1

シーケンスID (23)	送信先ポート1 (24)	送信先ポート2 (25)
1	下位ポート1	下位ポート2
2	下位ポート7	下位ポート8
3	下位ポート3	下位ポート4
4	下位ポート5	下位ポート6
⋮	⋮	⋮

22

【図８】

RAID5(IO Generate)

シーケンスID (29)	送信先ポート1 (73)	送信先ポート2 (74)	送信先ポート3	送信先ポート4 (72)
1	下位ポート1	下位ポート2	下位ポート3	下位ポート4
2	下位ポート3		下位ポート1	下位ポート4
3	下位ポート4		下位ポート1	下位ポート3
4	下位ポート5	下位ポート6	下位ポート7	下位ポート8
5	下位ポート7		下位ポート5	下位ポート8
6	下位ポート8		下位ポート5	下位ポート7
7	下位ポート9	下位ポート10	下位ポート11	下位ポート12
8	下位ポート11		下位ポート9	下位ポート12
9	下位ポート12		下位ポート9	下位ポート11
⋮	⋮	⋮	⋮	⋮

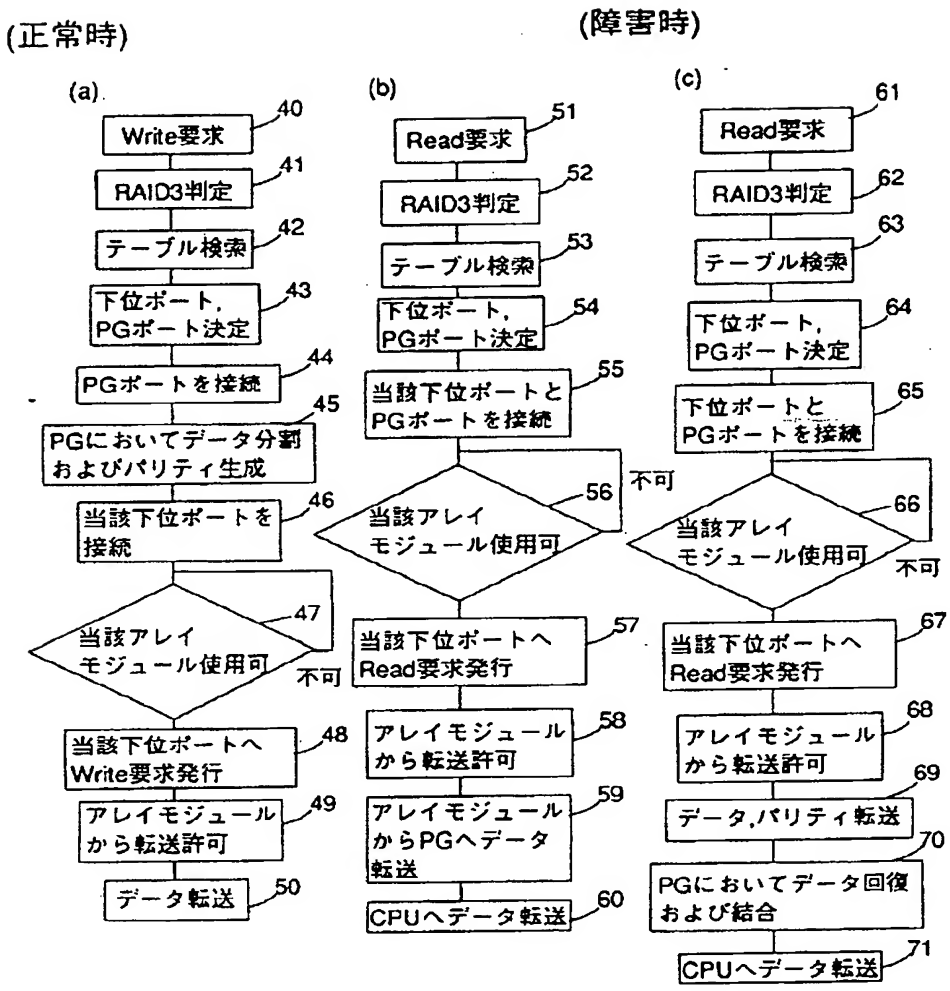
【図６】

RAID3

シーケンスID (23)	送信先ポート1	送信先ポート2	送信先ポート3	送信先ポート4	送信先ポート5
1	下位ポート1	下位ポート2	下位ポート3	下位ポート4	下位ポート5
2	下位ポート6	下位ポート7	下位ポート8	下位ポート9	下位ポート10
3	下位ポート1	下位ポート2	下位ポート3	下位ポート4	下位ポート5
4	下位ポート6	下位ポート7	下位ポート8	下位ポート9	下位ポート10
⋮	⋮	⋮	⋮	⋮	⋮

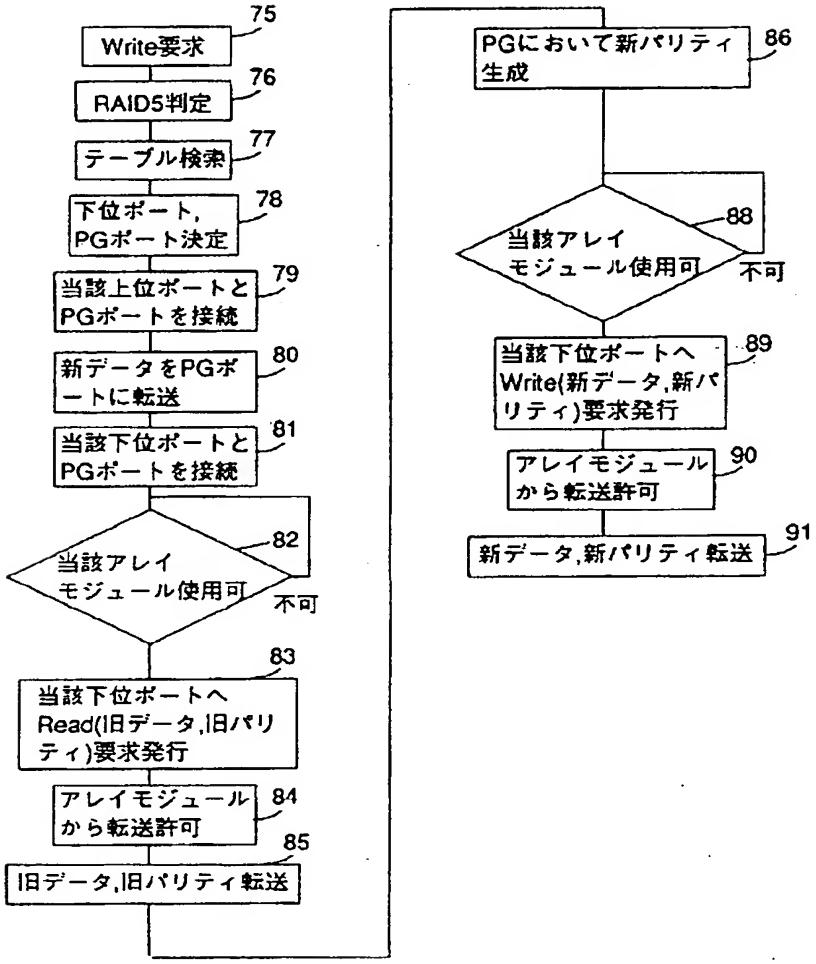
35

【図7】

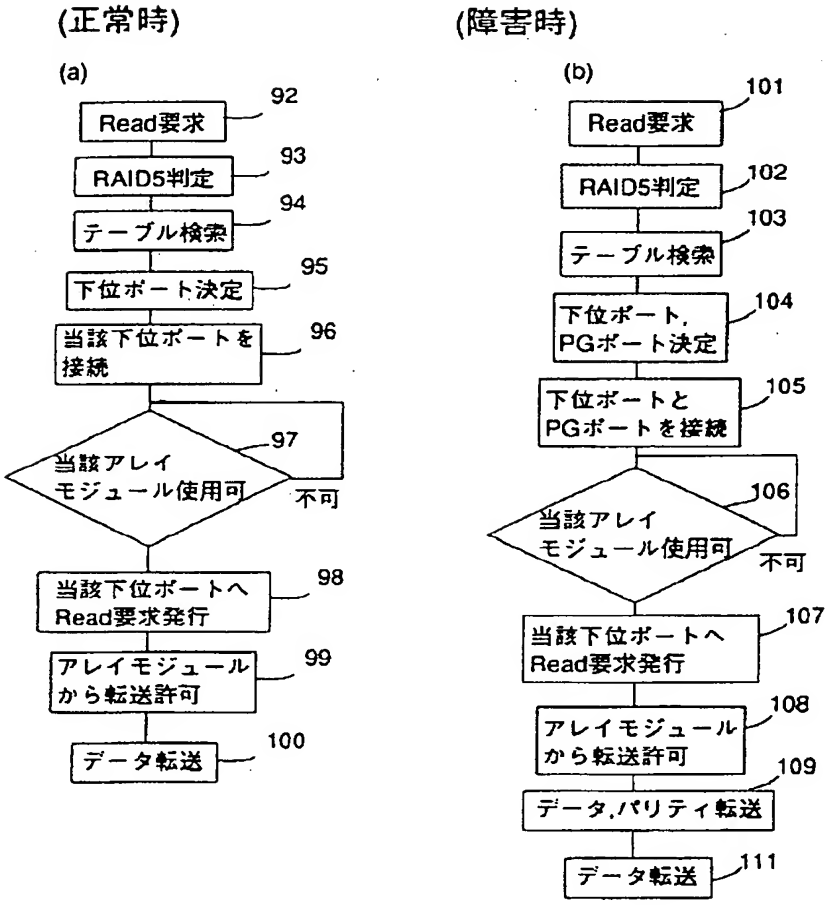

(正常時)
(障害時)
(a)
Write要求
40
RAID3判定
41
テーブル検索
42
下位ポート，PGポート決定
43
PGポートを接続
44
PGにおいてデータ分割およびパリティ生成
45
当該下位ポートを接続
46
当該アレイモジュール使用可
47
不可
当該下位ポートへWrite要求発行
48
アレイモジュールから転送許可
49
データ転送
50
(b)
Read要求
51
RAID3判定
52
テーブル検索
53
下位ポート，PGポート決定
54
当該下位ポートとPGポートを接続
55
当該アレイモジュール使用可
56
不可
当該下位ポートへRead要求発行
57
アレイモジュールから転送許可
58
アレイモジュールからPGへデータ転送
59
CPUへデータ転送
60
(c)
Read要求
61
RAID3判定
62
テーブル検索
63
下位ポート，PGポート決定
64
下位ポートとPGポートを接続
65
当該アレイモジュール使用可
66
不可
当該下位ポートへRead要求発行
67
アレイモジュールから転送許可
68
データ,パリティ転送
69
PGにおいてデータ回復および結合
70
CPUへデータ転送
71

【図９】

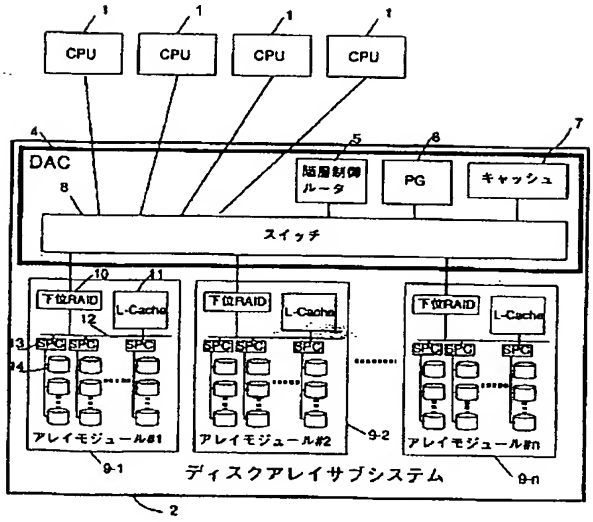
(正常時)


75
Write要求
76
RAID5判定
77
テーブル検索
78
下位ポート,
PGポート決定
79
当該上位ポートと
PGポートを接続
80
新データをPGポートに転送
81
当該下位ポートと
PGポートを接続
82
当該アレイ
モジュール使用可
不可
83
当該下位ポートへ
Read(旧データ,旧パリティ)要求発行
84
アレイモジュール
から転送許可
85
旧データ,旧パリティ転送
86
PGにおいて新パリティ
生成
88
当該アレイ
モジュール使用可
不可
89
当該下位ポートへ
Write(新データ,新パリティ)要求発行
90
アレイモジュール
から転送許可
91
新データ,新パリティ転送

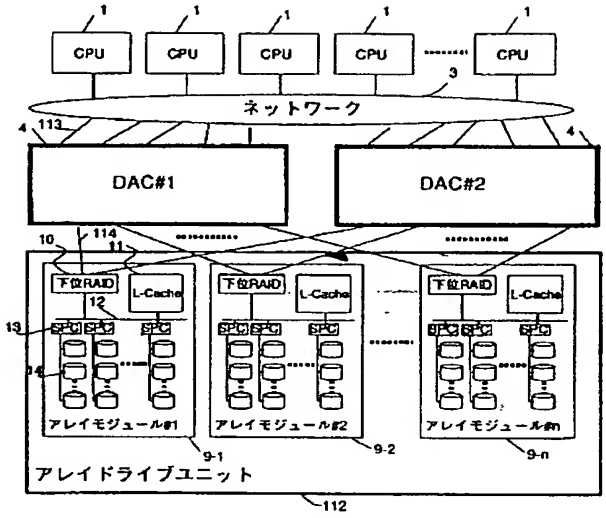
【図１０】


(正常時)
(障害時)
(a)
Read要求
92
RAID5判定
93
テーブル検索
94
下位ポート決定
95
当該下位ポートを接続
96
当該アレイモジュール使用可
97
不可
当該下位ポートへRead要求発行
98
アレイモジュールから転送許可
99
データ転送
100
(b)
Read要求
101
RAID5判定
102
テーブル検索
103
下位ポート,PGポート決定
104
下位ポートとPGポートを接続
105
当該アレイモジュール使用可
106
不可
当該下位ポートへRead要求発行
107
アレイモジュールから転送許可
108
データ,パリティ転送
109
データ転送
111

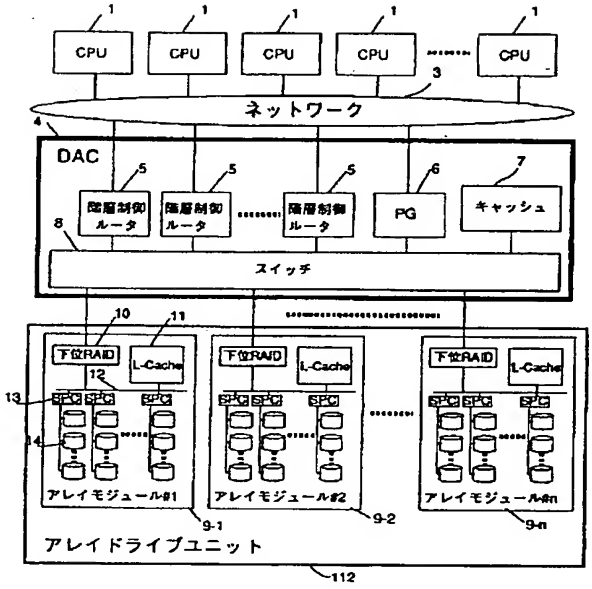
【図１１】


CPU
DAC
スイッチ
階層制御ルータ
PG
キャッシュ
下位RAID
L-Cache
SPC
アレイモジュール#1
アレイモジュール#2
アレイモジュール#n
ディスクアレイサブシステム

【図１３】


CPU
ネットワーク
DAC#1
DAC#2
下位RAID
L-Cache
SPC
アレイモジュール#1
アレイモジュール#2
アレイモジュール#n
アレイドライブユニット

【図１２】



フロントページの続き

(72)発明者　加茂　善久
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内